DOCUMENT RESUME

ED 387 497 TM 023 600

AUTHOR Taylor, Catherine S.; Nolen, Susan Bobbitt
TITLE A Question of Validity: A Model for Teacher Training in Assessment.
PUB DATE Apr 95
NOTE 52p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS College Faculty; Course Evaluation; *Educational Assessment; Education Majors; Higher Education; Instructional Effectiveness; Models; *Portfolios (Background Materials); Reliability; Self Evaluation (Individuals); *Teacher Education; Teaching Methods; Test Construction; Training; *Validity
IDENTIFIERS *Preservice Teachers

ABSTRACT

An assessment course for teachers is described that uses validity as the primary focus. Five dimensions of validity for classroom-based assessments are highlighted in a course that relies on the use of a process portfolio. The five dimensions are: (1) looking at the content of the assessment in relation to the content of the domain of reference; (2) probing the ways in which individuals respond to the items or tasks and examining the relationships among responses to the tasks and items; (3) investigating differences in assessment processes and structures over time, across groups and settings, in response to instructional interventions; (4) surveying relationships between assessments and other measures or background variables; and (5) tracing the social consequences of interpreting and using test scores. Excerpts from the self-evaluation of 1 cohort of 27 preservice teachers are presented to show the substantive nature of their learning about validity and reliability that resulted from using this instructional model. Results are also presented from studies comparing students who had more traditional assessment courses with those who were in the portfolio course. Study 1 compared course evaluations across teaching faculty for the two versions of the course for 8 instructors for the traditional course and 3 for the portfolio course. Study 2 involved analyses of data from surveys completed by 73 teacher education students, and Study 3 examined exit surveys from 280 students in the teacher education program. Results of all three studies supported the relevance and usefulness of the portfolio course and its focus on validity. Ten tables present study data. (Contains 24 references.) (SLD)

A Question of Validity: A Model for Teacher Training in Assessment

Catherine S. Taylor and Susan Bobbitt Nolen

University of Washington

Paper presented at the annual meeting of the National Council on Measurement in Education,

San Francisco, April 1995

More than ever before, pressure is being placed on teachers to create high quality assessments of their students' learning. Work is now underway in Kentucky, Vermont, New Mexico, and in the nineteen states that are members of the New Standards Project (Resnick and Resnick, 1991) to explore the viability of classroom-based projects and portfolios as sources of state or national accountability data about student learning. Teachers' ability to develop appropriate classroom-based assessments is now seen as one of the six core functions of teachers (Gulickson, 1986). Stiggins (1992) has noted that teachers spend at least one-third of their instructional time engaged in assessment-related activities. At the same time, research has shown that few teacher preparation programs provide adequate training in assessment (Schafer & Lissitz, 1987, Stiggins & Bridgeford, 1988), teachers do not believe they have the training needed to meet the demands of classroom assessment (Wise, Lukin, & Roos, 1991), teacher-developed assessments largely focus on testing for facts and simple applications despite the fact that teachers are attempting to teach higher order thinking and problem-solving skills (Stiggins, 1988), and teachers use published tests and supplemental materials without adequate skills in evaluating the appropriateness of these materials (Airasian, 1991, Stiggins, 1991).

A careful review of the most recently published beginning measurement text books shows that few are actually written with the needs and realities of teachers in mind. Most focus on classical measurement principles, interpretation of standardized test scores, methods for establishing estimates of reliability and validity, methods for obtaining item analysis information, and techniques for creating end of unit or course tests of student learning. The classroom reality, on the other hand, is one of fairly constant formal and informal assessment through observation, homework and in-class assignments, and an increasing use of alternative methods of assessment (Airasian, 1993; Stiggins, Faires-Conklin, & Bridgeford, 1986). A growing body of research (e.g., Schafer & Lissitz, 1987; Stiggins & Faires-Conklin, 1988) suggests that teachers do not perceive the information learned in traditional assessment courses to be relevant to their tasks as classroom teachers. There is a mismatch between topics covered and teachers' perceptions of their assessment needs (Linn, 1990). A mechanism is needed to bridge the gap between what teachers

are receiving during teacher preparation programs and what teachers actually need in the classroom while still helping them understanding fundamental assessment concepts and issues.

A number of authors have outlined what they believe are the essential understandings about assessment teachers must have in order to confront the ongoing assessment demands in the typical classroom (Airasian, 1991; Linn, 1990; Stiggins, 1991). Each of these authors highlights the need for teachers to create assessments that are appropriate for the instructional methods and subject matter foci in a given classroom. This focus on instructional context can easily be lost when approaching the sizable task of teaching teachers about assessment. In teacher education programs, there is often a single assessment course for all prospective teachers, from the kindergarten teacher, to the APP calculus teacher, to the middle school vocal music teacher. In response to the formidable range of assessment content teachers need to know, instructors may design survey courses that result more in intellectual awareness than actual competency. In addition, pre-service and in-service teachers have very strong personal theories about classroom assessment, based on their own experiences as students, as student teachers, and as teachers.

Creating a new and more appropriate understanding about classroom assessments can be a difficult task. However, in building a strong conceptual understanding of the relationship between classroom assessment practices, content-area disciplines, and instructional methods, teachers begin to understand the place of assessment in the classroom. These connections lay the foundation for the validity of classroom-based assessments.

In this paper, we describe an assessment course for teachers that uses validity as the primary focus and we highlight five dimensions of validity for classroom-based assessments. Excerpts from the self-evaluations from one cohort of pre-service teachers is presented to show the substantive nature of their learning about validity and reliability, the cornerstones of assessment, that resulted from using this instructional model. In addition, results from three studies are presented. These studies compared students who had a more traditional assessment course with those who were in the portfolio course in terms of course evaluations, perceptions of the course via exit surveys, and follow-up transfer of the learning in the course to their field experiences.

Background

The assessment course was taught at a large, northwestern university that provides a certification program for approximately 250 elementary and secondary teachers per year. The university is on a ten-week quarter system and a given class included pre-service teachers from all academic subjects and the arts for kindergarten through twelfth grade. During the quarter in which the assessment course was taught, students spent at least 20 hours per week in their field placement sites in addition to their course work as a transition into full time student teaching the following quarter.

During the summer of 1991, the focus of the traditional tests and measurement course was redesigned. Prior to this time, the course had been taught in a fairly traditional manner (see Gullickson, 1986; Schafer & Lissitz, 1987; Stiggins & Faires-Conklin, 1988). The course covered standardized test interpretation, item writing and item analysis techniques, and statistical procedures for obtaining estimates of validity and reliability of tests. Students were assessed on their ability to write test items in various formats, and tested on their knowledge of measurement principles and concepts. The decision to redesign the course was based on research suggesting that teachers do not benefit from traditional measurement courses and do not see these courses as having value for their work as teachers (Schafer & Lissitz, 1987, Stiggins & Faires-Conklin, 1988) as well as recommendations about what assessment courses for teachers should address (Airasian, 1991; Linn, 1990; Stiggins, 1991).

In redesigning the course, the two most significant shifts were that a) all assessment concepts were to be taught in the context of instructional practices and b) the major emphasis of the course was to be on assessment validity rather than simply assessment techniques. We began with a model proposed by Linn (1990), and expanded it to include the use of process portfolios (Valencia, 1990; Wolf, 1991). The resulting course resulted in significant learning about assessment strategies and concepts. Every effort was made to model feasible and effective use of portfolios in teaching and as well as a variety of other assessment practices.

The framework for the dimensions of validity for classroom-based assessments was

derived from the work of Messick (1989). Messick views construct validity as the core issue in assessment, and states that all inferences based upon, and uses of, assessment information require evidence that supports the inferences drawn between test performance and the construct an assessment is intended to measure. In his discussion of validity, Messick discussed various paradigms within the philosophy of science and what these paradigms suggest as the appropriate methods for obtaining evidence of validity for assessments. He indicated that there were "only a half dozen or so" distinct sources of evidence for the validity of assessments (p. 16).

> We can look at the content of the test in relation to the content of the domain of reference. We can probe the ways in which individuals respond to the items or tasks. We can examine the relationships among responses to the tasks, items, or parts of the test, that is, the internal structure of test responses. We can survey relationships of test scores with other measures and background variables, that is, the test's external structure. We can investigate differences in these test processes and structures over time, across groups and settings, and in response to . . . interventions such as instructional . . . treatment and manipulation of content, task requirements, or motivational conditions. Finally, we can trace the social consequences of interpreting and using test scores in particular ways, scrutinizing not only the intended outcomes, but also the unintended side effects. (p. 16)

Messick stated that multiple sources of evidence are needed to investigate the validity of assessments. In the classroom context, this means that teachers must know how to look at their own assessments and assessment plans for evidence of their validity and they must know where to look for alternative explanations of student performance. The five dimensions of validity evidence teachers can consider include the following:

Dimension 1: Looking at the content of the assessment in relation to the content of the domain of reference. Before teachers can look at their assessments in this way, they must be able to think clearly about their disciplines, understanding both the substantive structure (critical knowledge and concepts) and the syntactic structure (essential processes) of the disciplines they teach. They must be able to determine which concepts and processes are most important and which

are least important in order to adequately reflect the breadth and depth of the discipline in their teaching and assessments. As Messick (1989) states, one of the greatest sources of construct invalidity is under-representation of some dimension of the construct. Once they have clearly conceptualized the disciplines they teach, teachers must know *how* to ascertain the degree to which the types of assessment tasks used in the classroom are representative of the range and relative importance of the concepts, skills, and thinking characteristic of subject disciplines. Teachers must also be able to answer the question, "Do the assessments give me information about all of the targeted concepts and processes?"

In addition, because the process of assessment is as much a function of how performances are scored as it is a function of whether the tasks elicit student learning related to the structure of the discipline. Teachers must examine the degree to which the rules for scoring assessments and strategies for summarizing grades reflect the targeted learnings. As with breadth and depth of coverage within assessments, teachers must evaluate whether scoring rules give too little or too much value to certain skills leading to questions about the validity of the interpretations teachers make from resulting scores.

Dimension 2: Probing the ways in which individuals respond to the items or tasks and examining the relationships among responses to the tasks and items. Teachers must examine the degree to which the assessments actually elicit the learning students are to achieve in a subject area. They must know how to examine their assessments for their potential to draw out the targeted learning. This means they must examine the tasks given to students to determine whether students are really being asked to *show* the learning related to the targets and they must use assessment strategies that will allow them to probe their students' thinking and processes. This becomes increasingly important as higher level thinking and processes are considered important learning targets and must then be assessed. In performance assessments, for example, examinees are often asked to explain their thinking and reasoning as part of the assessment task. Teachers commonly ask students to show their work in mathematics and science assessments. These classroom assessment practices lend themselves well to probing the ways in which individuals are

responding. In general, teachers must ask themselves, "Do the assessment tasks effectively elicit the targeted concepts and processes? Are some assessment formats more effective in eliciting student learning related to a given aspect of the discipline than others?"

Teachers must be able to look across student responses to a variety of assessment tasks to determine whether patterns of student responses support the use of the assessments. The mechanisms for this type of examination have historically been quantitative item analysis techniques. Research has shown that few teachers use these quantitative techniques in actual classroom practice (Stiggins & Faires-Conklin, 1988). Teachers, however, can be shown how to scrutinize student work qualitatively, looking for patterns in their responses that reveal positive and negative information about the validity of the assessments. They must know how to look at student responses across a range of items and tasks and ask themselves, "Did I actually teach these concepts well enough for students to perform well? Have the directions for the task or the wording of the items limited my students' understanding of the expectations of the task? Are students who show understanding of a concept in one assessment format (e.g., an essay), also showing equal understanding in a different format (e.g., a multiple-choice test)?"

Dimension 3: Investigating differences in assessment processes and structures over time, across groups and settings, in response to instructional interventions. To investigate these validity issues, teachers must know how to examine the relationship between the instructional practices used and the assessments themselves. They must also evaluate the adequacy of various assessment strategies for the unique needs of their students. Good teachers alter instruction over time and for different groups. They must be able to judge whether an assessment can be used in many different contexts or whether varied contexts, groups, and instructional strategies require the development of different assessments. They must be able to judge whether the assessment strategies they plan to use are appropriate for the groups they teach and consistent with instructional methods they use.

Dimension 4: Surveying relationships between assessments and other measures or background variables. Teachers must know how to ascertain the degree to which performance on the assessment and the score resulting from the assessment are directly attributable to the targeted

learning. They must determine whether performance is influenced by factors irrelevant to the targeted learning such as assessment format, response mode, gender, or language of origin. This becomes increasingly critical as classrooms become more diverse and whole group teaching becomes more difficult. For example, if a teacher has students who are poor readers, she or he must ask, "Is it appropriate to place a heavy reading load on a mathematics or science task designed to show students' understanding of mathematics or science concepts. Can the task be adapted to meet the needs of the poor readers." In general terms, teachers must know how to adapt an assessment format to meet the needs of diverse students while still obtaining good evidence about student learning. Teachers must ask themselves, "Does the use of one assessment format for all students lead to bias in the assessment process?" Finally, teachers must know how to create scoring mechanisms for open-ended performances that are clearly related to the learning targets and that are precise enough to prevent biased scoring. Teachers must ask themselves, "Have I created an assessment task and scoring mechanism that allows me to insert my own biases about students as I evaluate their work?"

Dimension 5: Tracing the social consequences of interpreting and using test scores in particular ways, scrutinizing not only the intended outcomes, but also the unintended side effects. Teachers must consider the influence of classroom assessments on the learners themselves. The nature of the assessments, feedback, and grading can all influence student learning, students' self concepts and motivation, and their perceptions of the disciplines being taught. Assessments also show students what is most valued in a subject area. Teachers who assess their students' knowledge of science by giving them multiple-choice tests of isolated facts, for example, communicate that science is a collection of facts on which everyone agrees. Those who assess students' inquiry strategies and their ability to make generalizations from observations or to systematically test their own hypotheses, communicate something different about the structure of the discipline of science.

The Structure of the Assessment Course

The dimensions of validity served as the focus of the course. To address all five

dimensions, it was necessary to help students investigate assessment concepts in a meaningful context. A process portfolio provided both the means for instruction and learning during the course, and the product used to assess students' learning at the end of the course. If our students were to see assessment as a powerful teaching tool (Wolf, 1990), we needed to provide an opportunity for students themselves to experience assessment as a powerful learning tool. The use of process portfolios allowed students to benefit from peer and teacher feedback on the first draft of each component prior to its submission for grading purposes. Instructor feedback was intended to guide their learning so that subsequent versions of their work reflected a better understanding of the course objectives. Feedback in this case was not editorial but was designed to lead them to clarify their own thinking and thereby improve the quality of their own work.

In order to allow students to see the connection between assessment and instruction, the heart of the portfolio was an integrated unit plan. Students outlined a term-long plan for a subject they would be likely to teach, and produced planning documents that were authentic representations of the type of work teachers do. These plans included subject area goals and objectives, a plan for a 2-4 week unit from that subject, a grading policy, and unit assessments. Students were required to write rationales or justifications for all assessment decisions made during the development of components of the plan. Finally, students wrote self-assessments of their understanding of major assessment concepts for the course.

The components listed above formed the core assignments of the course as it evolved over the next twelve quarters. Based on student work and feedback, we adjusted the portfolio components and instructions and experimented with various scoring schemes for the final portfolios. In what follows, we describe each of the portfolio components in some detail and in the order the components were assigned. We then discuss the link between each component and the validity framework using the students' own writing to illustrate their thinking about these validity issues. The excerpts come from pre-service teachers who were enrolled in the course during the spring of 1994. Excerpts from all 27 students' work appear in the following pages. Identification numbers are given with each excerpt. In the self-evaluation they were required to:

1. Discuss their current understanding of the concepts of validity, reliability, bias, and fairness as well as what specific work in the course had helped them understand these concepts and why.

2. Select any eight of the sixteen assessment course objectives and discuss what they had learned related to that objective, what aspect of the course had helped them to learn it, and how.

3. Select one component of the portfolio and describe how the submit-revise-resubmit cycle had influenced their learning.

4. Discuss their conceptions of assessment prior to the course and how these conceptions had changed during the course (if conceptions had not changed they were to discuss why).

Course description, goals and objectives

Students began by writing a brief description of a subject area they planned to teach. The description included a general outline for one quarter or trimester, including the major concepts and processes to be taught during that period of time. Students were encouraged to work with their cooperating teachers to frame a description for a subject that they might actually teach during their full time student teaching experience the following quarter. Some students, however, chose to develop a portfolio for a subject they *hoped* to teach. As one student put it, "I developed this portfolio so that I could keep my hopes up for the kind of teacher I want to be rather than the kind of teacher my cooperating teacher will let me be."

Once the general description was completed, students wrote goals and learning objectives for the subject area. While goals and objectives are often an aspect of an assessment course, we broke from assessment tradition by having students write objectives in terms of what their students would *learn* rather than what they would *do to show their learning* (e.g., "Students will learn how authors develop themes in literature" rather than "Students will identify the three major themes in the novel Ana Karanina."). Students had access to various standards documents (e.g., National Council for Teachers of Mathematics (NCTM) Curriculum Standards) as they wrote their objectives, but were encouraged to draw from their own values and beliefs as well. We hoped that this level of objective writing would lead our students to clarify, for themselves, the most central

learnings in their disciplines. This conceptual clarity is necessary if teachers are to develop assessments that reflect the disciplines studied (Validity Dimension 1).

Finally, students wrote a rationale describing how their goals and objectives reflected the substantive and syntactic structures of the disciplines they intended to teach. This requirement built upon the educational psychology course they had taken the previous quarter in which they explored the concepts of disciplinary structure (Schwab, 1978) and pedagogical content knowledge (Grossman et al., 1989).

As students struggled with how to make the structures of their disciplines real for themselves (thus defining their constructs (Messick, 1989)), they found themselves revising objectives and goals. They discovered that the standards provided by national and state documents were often vague, making it difficult to understand what was really meant. They also discovered that different students developed different arrays of goals and objectives for the same disciplines, reflecting different notions of discipline structure. This is in keeping with Messick's idea that constructs are theorists' inventions and that there may be differing definitions of constructs. Students struggled through this iterative process throughout the quarter as they found that some objectives that seemed to represent the discipline were too fuzzy or two narrow when it was time to design instruction and assessments.

In their final self-evaluations, many students indicated that this aspect of their portfolio work was the most significant; they felt it helped them become more grounded in their subject areas. The following excerpts from students' self-evaluations highlight this point.

> "Writing the goals and objectives rationale was particularly challenging. Having to think about the structure of my discipline and then relate the objectives that I think are important to this structure forced me to think about how I will teach chemistry as well as why I will teach chemistry." (6)
>
> "The subject area [component] was able to teach me something about validity. When I decided to teach jazz improvisation, I did so because I love to improvise jazz as a professional musician. My dilemma was, what are the most important

aspects of jazz improvisation that I can teach to high school students so that they will be able to improvise jazz in an adequate to more than adequate level? I had to decide what the most important and fundamental concepts of jazz improvisation were. I have spent so much time going beyond the basics, studying the intricacies of jazz improv. that I wasn't sure what the most valuable and valid concepts are to teach. By making myself spell out what I consider the most fundamental building blocks of jazz improvisation, I had created a working unit plan." (2)

"I think that the first step in learning how to ensure a direct relationship between objectives and assessment, and learning the appropriate relationship between instruction and assessment, is to articulate goals and objectives for my discipline. If I do not know what is essential in Language Arts, I cannot adequately instruct or assess the learning." (4)

Although they did not find the experience enjoyable, they realized the long term benefits of this type of thinking, not only for the course they intended to teach, but for other courses they would probably teach.

"The best part of the course for me was the subject area description and goals because it forced me to stop and think about why I want to teach biology. . . .Being a good teacher is a difficult task. The best way to overcome this is going through the process we went through during the development of subject description, goals, objectives, and rationale. I feel that an entire course should be developed covering this. . . . It will help me down the road as a teacher." (7)

". . . thinking about grading and assessment does not begin after teaching with the creation of a test. It starts with the thinking that you do when you are first creating a unit, and it is an extremely important part of the planning you do. When you ask yourself 'What do I really want them to learn?' and when you develop rationales for teaching these things, you are taking the first critical step towards developing a unit that displays a high degree of internal consistency, which increases the validity of

your unit. This type of consistency is very difficult to achieve if you have no clear goals and objectives or ideas about why they are important. . . . The subject area description, goals and objectives are a powerful exercise in self communication that brings clarity to the thoughts that are the original source of inspiration for a unit." (22)

"Before writing this component I felt goals and objectives would be 'handed down' by the school board, principal, etc. Much like orders are given in the Army (10+ years experience). Since a text is provided I would know what to teach, but this assignment and student teaching shows me there are many teacher decisions to be made on goals and objectives, not just how to teach today's lesson." (19)

"Without learning objectives, a teacher would be lost in teaching and assessment. One must first think about what it is you really want to 'teach' before doing so and then assessing student learning. By focusing my thoughts, my teaching becomes more fluid and has its own focus, and my assessments become more valid. Without learning objectives you become like a 'leaf in the wind,' struggling to do something that feels right as a teacher." (15)

"From what I had learned about goals and objectives before, I thought they were just busy work that had to be done for the Education program. How foolish I was! I see now how important it is for both me and my students to know and understand what it is I am teaching and they are to be learning. Goals and objectives are the basis for everything that is done in a unit, and they should never be taken lightly. I see now that if these are done right and that if a lot of thought and effort are put into them, the rest of the unit falls nicely into place." (8)

"I learned that I should begin my teaching process with my learning objectives. I need to make it clear what I want my students to learn. Then I need to decide how I am going to assess these objectives. . .then I can plan activities that will tailor the student learning towards both the objective learning as well as the form of

assessment. It's like looking at a boulder, envisioning the final work of art, setting the parameters for how it will be judged, and then finally picking the tools and using them on the boulder." (2)

"I am very good at doing things off-the-cuff. This course has forced me to sit back and take a look at what are the important learnings in my discipline, and to make that importance the criteria for leading my students in an activity, not just an afterthought. . . . I really had to teach myself to start with the essential learnings and build activities, rather than the other way around." (3)

"I learned through doing the description paragraph and the rationale section, how important it is to know the why's behind what you're teaching - both to clarify things in your own mind and to be able to answer potential questions from students, parents, and administrators." (9)

"I see that developing appropriate objectives, activities, and assessments is a dialectical, iterative process - it is not as if objectives can always be thought of first in a linear fashion - they also well up from the process of teaching and assessment. All 3 should be responsive to each other and change if necessary. This is a dynamic process." (20)

Finally, students used this clarification of their disciplines as the basis for deciding what types of assessment information were most valuable in helping them ascertain whether their students were, indeed, achieving these targets.

"It first made me focus on what I really wanted my students to learn, and then I had to find different and appropriate ways to assess whether or not the students learned these things. If one of my unit objectives was to view the American Revolution and its effects from a variety of perspectives, then an assessment that only deals with one perspective is not a valid assessment. It does not tell me if they have learned what it is I want them to learn." (15)

"I felt that my performance assessment was relevant to the discipline of

mathematics, the targeted objectives, and what I taught in class. Therefore, I felt that the information I received from the student performances was more reliable than if they had been given the traditional item set. . . . I wouldn't want to use my traditional item set to assess how students connect mathematics to the real world. The item set is not designed for that purpose. Using the appropriate format to assess specific kinds of objectives increases the validity of the assessment." (5)

"I realized that traditional items are probably the easiest way to assess and keep personal feelings aside, however, this type of assessment does not stress communication and supporting evidence which would be more true to my discipline." (9)

"This component [performance assessment] taught me that performances are an extremely important form of assessment. Performances are valid if they fit with the structure of the discipline and relate to essential learning objectives. . . . Performances are especially important because they ask students to do the things that 'real' people do in the discipline." (4)

"In music it is a challenge to elicit individual behavior when the general course objective is ensemble oriented. I enjoyed having to think about actual behaviors that were observable and how that observation would practically be handled in a group. Checklists allow this to happen without a big deal being made of the process." (23)

Unit description

Once students had completed their subject area descriptions, they selected a two to four week unit of study as the focus of the remainder of the portfolio components. Students who were in the arts had more difficulty with the idea of arbitrarily cutting a their subject area into two to four week segments and usually focused instead on a single dimension of a fairly long-term, integrated course of study (e.g., music teachers might focus on music theory fundamentals for the entire nine to twelve week term even though their students would also be learning performance skills and music history simultaneously). Again, students were encouraged to focus on a unit they would

actually teach. While this component of the portfolio is not typically included in an assessment course, it proved vital to students' understanding of how to establish the validity of assessments. Without the instructional unit as an anchor, it would be difficult to address aspects like the validity of methods of assessment for the methods of teaching used (Validity Dimension 3).

Students selected up to six subject area objectives as the focus for the instructional unit. This proved to be another sticking point. Faced with textbooks that listed as many as ten objectives for a single day's reading, they found it hard to imagine selecting only six objectives as the focus of a two to four week period. Students often listed more than six objectives in the first drafts of their unit descriptions. However, students discovered that objectives written at the level of major disciplinary understandings, rather than as narrow behaviors, take time to teach and adequately assess (Validity Dimension 1). In the end, many students pared their objectives down to four or five in the final draft of the unit description.

> "When I revised my instructional unit assignment, the first thing I did was remove 1 of my objectives having to do with the differences between French and American teenage culture. . .I wasn't really teaching it, so couldn't really assess it." (9)
>
> "The initial revision helped make me aware of the danger of covering too much subject matter superficially. I tried to include activities in each part of the unit that would enrich the experience of the musical work, but I wonder if students would focus on the details rather than the whole of the musical work." (23)
>
> "I must limit the scope of what I will try to achieve within a unit . . . the objectives for a unit must be compatible. This seems obvious, but in writing my first draft, I got so caught up in deciding what I wanted students to learn that I forgot how I would teach the objectives, as well as how the objectives would (or would not) work together." (10)
>
> "I tend to be very traditional, having gone through a traditional curriculum myself in high school in the 60's. I felt the text would provide objectives, activities (pretty much limited to homework and tests!) and that justification was that a higher

authority decided these. I need to look harder at objectives for each unit, and how each lesson reaches these objectives." (19)

In addition to selecting objectives, they wrote a day by day narrative of the activities they would use to teach the objectives, linking the objectives to each activity, and providing a rationale for why the given activity or activities would lead to the targeted learning. Students were told to "create a narrative 'video-tape' of the unit as if they were telling a friend what they and their students did each day" rather than discussing the content taught. Again, although this work was difficult, students indicated that they benefited from having to justify each activity in terms of how it would lead to the targeted learning. They found they were assessing the validity of their instructional activities for the learning objectives they had targeted. This evaluation gave students more confidence when making decisions about the fit of assessments to the targeted learnings and the methods of instruction (Validity Dimensions 1 and 3)

> "I learned the importance of planning teaching strategies to provide students with the opportunity to think about and do precisely those things which I have said that I want them to learn. This fit between targeted objectives and teaching strategies, which is such a critical precursor of valid assessment, does not occur spontaneously. You have to think hard about whether the activities you have created and the assignments you give will teach the intended learning. If they do not, and yet you assess for student learning of these objectives, your assessments will be invalid. Doing the justifications for the unit activities helped me understand the importance of thinking about how activities teach objectives and why I needed to teach in a certain way, given a particular objective. If I have thought critically about this 'how and why,' I can share the information with students, which will make learning activities more meaningful, and assessments fairer." (22)
>
> "In drafting the unit activities description, I did not explain <u>how</u> I would teach the students to read critically and articulate their own interpretations. I was forced to examine HOW I planned to teach students these skills, which I had always taken

for granted as abilities which students just 'pick up.' . . . I realized that it would not be fair to assess students' ability to develop their own interpretations if I had never taught them how to do it. I discovered that saying, "they will learn this through discussion," is an inadequate description of how I will teach a skill" (10)

"In both developing and revising this component I became very aware of the relationship between objectives and instructional activities. I feel that I am creative and have enough energy and inspiration to do the good work of teaching. Yet, the teachings of this component will help focus my energies. I see the useful need of designing classroom activities that teach my pre set objectives. That way, my classes can have a clear process path, like a good story, and not flounder around toward an amorphous end, like a bad story." (18)

"The component of the portfolio that I think influenced my learning the most about the relationship between targeted objectives and instruction was the unit plan. It was extremely difficult to precisely match my instruction to the targeted objectives I had written. In turn, it was even more difficult to write justifications for activities that did not always teach the targeted objectives. I really struggled with this component. . . . I learned that objectives that seem to be appropriate learning targets are often very difficult to teach within the framework of a unit. I also learned that it helps to have an idea of the way in which objectives will be taught when developing them. . . . If learning objectives are strong, activities . . . are not nearly as difficult to plan. This assignment has already affected my teaching; when planning the unit I actually taught in my placement, I carefully thought out how each activity I planned related to my objectives." (6)

"When I first began creating my unit activities component draft, validity didn't mean anything to me. I was too focused on putting together interesting and creative lessons. . . . What I first turned in was a conglomeration of lesson plans that really weren't connected to each other, let alone connected to the learning objectives. . . .

[W]hen I received my first draft and began to revise it I realized that my lesson plans did not coincide with my learning objectives at all. There was no focus and no underlying goal. . . . I was just grasping at straws. I really didn't understand how the underlying structure of a unit needed to be connected and based off of strong, concrete objectives. I thought that all you needed was a few good lesson plans - now I see how it all connects. . . . However, after . . . seeing how often I asked students to be responsible for material I never had even taught them I finally understood the connection between the learning objectives, teaching, and assessment." (13)

"Trying to plan activities that would effectively teach each of my objectives was very difficult, I realized that I could not fairly and validly assess my students if I did not plan activities that would teach them the targeted learnings. . . . I don't think I realized how important it is to plan activities that clearly and effectively teach learning objectives. Assessment is then the ultimate measure of the teaching as well as student learning." (6)

"At the conclusion of this course, I realize that it has become so obvious a point, to test on what you have taught. But when creating the assessment options for my portfolio I found that it wasn't as easy as it sounded. I found myself assuming that my students would have "gotten" my meaning even if I hadn't clearly taught it." (11)

Unit Assessments

For the next part of the portfolio, students used a variety of techniques to create assessments for their instructional units. They wrote a description of all the different assessments they planned to use in the course of the unit. These included pre-assessments (those designed to find out what their students already knew prior to teaching the unit), informal assessments (those that would be collected as part of the on-going instructional process and would be used to help them monitor their students' learning), and formal assessments (those that would be carefully

structured so that the teacher could obtain systematic evidence about student learning).

Students then fully developed four different types of assessment for their units: observational checklists or rating scales, performance-based assessments, essay items, and traditional items (multiple-choice, true-false, short-answer, completion, and matching items). Students were required to develop assessments that fit with their instructional methods and that assessed their unit objectives. Students then had to write a rationale for each item or task that answered several questions:

1. Will the item/task draw out their students' learning related to the unit objective(s) it is intended to measure?

2. Does the item/task reflect concepts, skills, processes that are essential to the discipline?

3. Does the item/task fit with the instructional methods used in the unit?

4. Do the rules for scoring the item/task relate directly to the unit objective(s) the item/task is intended to measure?

5. Is the mode of assessment such that all students who understand the concepts will be able to demonstrate them through the assessment?

These components of the portfolio gave students the most direct opportunity to explore all five dimensions of validity. By having to think about each item or task and its relationship to the discipline (Validity Dimension 1) and the unit methods (Validity Dimension 3), students were forced to go beyond simply practicing item or task writing techniques. While students often resisted the process of writing rationales, many stated in their self-evaluations that writing the rationales for the unit assessments helped them think about the kinds of assessments that would best measure the learnings they cared about.

> "By having to justify why a specific assessment was appropriate to the discipline, objectives, and teaching I created assessments and activities that directly followed my objectives and teaching. The justifications also helped me to understand the direct relationship between instruction and assessment. . . . Since I am more concerned that students develop their mathematical thinking skills than they

memorize formulas for a test, my instruction and assessment will reflect this." (5)

"Every time we wrote justifications we had to defend the validity of the assessment. Most of the assessments we studied are valid to a degree - depending on who and what they are intended to assess. But not every assessment is valid for every type of learning. The process of testing those assessments by justifying them and designing answers for them (as with the essays) became the tools for valid assessment." (14)

"The concept of validity . . . seemed straightforward and simple enough. It wasn't until I had to create assessments of my own, that I discovered how easy it is to create invalid assessments. . . . Writing rationales for the portfolio pieces kept the validity issue constantly before my eyes. Not only did it make me aware of when I was assessing something I hadn't taught. It also made me realize when certain unit activities or performance criteria or items were not effective at providing the kind of information I needed to make a valid assessment about student learning of the objectives I had targeted." (22)

Our students became more aware of the power of assessment to communicate to their students the nature of the disciplines, as well as communicating how the disciplinary learning relates to the world beyond school. They became aware that one potential consequence of assessment is a better or poorer understanding of the disciplines themselves (Validity Dimension 5).

"Assessments tell students what knowledge is most important within a discipline; if a student knows he/she will be required to memorize dates and names for a history exam, he/she will learn dates and names and not bother with conceptual understanding. Assessments drive students' learning, which is most obviously indicated by the plaintive, 'Is this going to be on the test?' If it's not, many students do not think 'it' is worth learning. Thus it is imperative that assessments be directly related to desired learning outcomes." (10)

"Assessments are not neutral! . . . Assessments send messages about a discipline; they communicate to students in a direct, concrete and powerful way about what is really important to know in this subject." (22)

Students had to address issues of bias and fairness in assessment when considering mode of assessment and its influence on performance for different groups. Could performance be attributed to some characteristic of the assessment or some irrelevant characteristic of their students rather than student learning (Validity Dimension 4)?

"There are those few [students], though, that I have trouble with wanting them to do well or to do poorly - and many times the papers end up leaning the way I feel about the student, instead of reflecting exactly what the students have done. . . . Well-defined criteria eliminate the temptation to apply the assessment criteria unequally, so I thought a lot about this issue while I was doing my checklists and rating scales. The presence of standardized criteria that are clear to me and to the students help guard against bias." (3)

"The students in my placement are intentionally given vague criteria. The teacher considers it her right to use her personal judgments of the student's attitude and behavior to influence the grade. If the criteria [are] not spelled out she has the leeway to insert her prejudice. Students realize what is going on and they become cynical and resigned. Few of them try to fight it. This lack of fairness is so wide-spread that they have come to expect it." (14)

"I began to think like a teacher as I revised my traditional items. I also learned how to use multiple choice and other traditional items to assess student knowledge of more than just trivial facts. I learned to use a more objective format to increase the diversity of the assessments in my unit. This will help me provide students who may not write or discuss very well, with an opportunity to demonstrate their learning, which will in turn reduce bias." (22)

Students also had to consider whether assessments were presented in a way that allowed

their students to demonstrate learning without guesswork, thereby enhancing the validity of the information they received from their students (Validity Dimension 2). Interestingly, students linked clarity of directions not only to validity (will the assessment elicit student learning related to the targets), but to *fairness* in assessment.

> "One of the most personally meaningful and exciting things I learned about assessment in the course is the power of sharing criteria, standards, and expectations with students, in other words, the power of being fair. . . . Making sure that assessments are fair, that students know exactly what is expected of them and what is being assessed as they complete an assessment activity, goes a long way toward helping students achieve desired learning outcomes. It is also the only ethical way to assessment and, almost inevitably, to grade. . . When assessments are fair, when criteria are made public, students do not have to read the teacher's minds. They can focus on doing exactly what they have been asked to do and they don't have to worry about any surprise criteria. . . " (22)
>
> "Giving the criteria for successful work helps make an assessment valid, as it insures that a student's essay demonstrates the student's conceptual and/or procedural understanding rather than his/her ability to read the teacher's mind. It is unfair to make students 'jump through hoops in the dark.' . . . Without careful planning, essay questions could become items which force students to 'spray and pray,' thus not assessing targeted objectives. The questions must be carefully structured in order to communicate to students what concepts should be discussed." (10)
>
> "When writing the Performance Assessment component . . . I was not originally clear enough about how students should demonstrate that they had learned how to use all of the systems in the library for gathering information. Because of the vague assessment instructions, I did not have enough information to confidently say that students had learned how to use all of the information systems. I modified the

assessment instructions and required them to document how they found information in the library in order to give me more information for making decisions about their learning." (12)

"Beforehand, my assessment of essays was largely global and intuitive. Developing and revising the performance criteria and scoring rubrics, however, helped me understand the importance of specifying the performance criteria if I am to be a good (fair) teacher. Without these criteria, students have little control over determining how well they perform. As a teacher, I want to treat students fairly and I want them to understand that the mechanisms for establishing validity and fairness can be scrutinized and understood - that their work and abilities are not simply being subjected to the arbitrary judgment of a thoughtless teacher. The performance criteria is there for all to see and work towards." (20)

"I . . . discovered that in order to assess students' learning of the targeted objectives, criteria and standards for a successful performance must be EXPLICIT. By clearly communicating to students what they must include in their paper to execute a successful performance, I give every student the opportunity for success. In addition, by outlining very specific and clear scoring rubrics, I lessen the chances of unfair grading practices on my part. As I just finished grading a thick stack of papers without an effective scoring rubric and am now suffering from occasional guilt pangs, I now know how important rubrics are for preserving the sanity of all parties involved!" (10)

"At the beginning of the year, I was designing essay questions that I thought were excellent, but were, in fact, poorly designed and unreliable. They could be interpreted many ways, and hence, students did not always show the learning I wanted them to." (15)

"Giving students clear expectations and then assessing those expectations increases the fairness of an assessment. Although I thought about expectations while creating

all four assessment [types], the performance assessment seemed to cement my understanding of the relationship between fairness and expectations. . . . I realized that when students understood expectations, they were able to concentrate on their learning." (5)

"Of all of the components of the portfolio, the performance assessment and essay items influenced my understanding of fairness the most. In developing both of these, I learned that it is very important to give clear directions to students that reflect both the learning objectives and the criteria that will be used to ultimately assess these students. . . . I also learned that it is possible to develop assessments that account for different learning styles." (6)

Students had to consider the appropriateness of a given assessment type for the unit objectives and activities (Validity Dimensions 1 and 3).

"As a longtime student, I have been frustrated with many of the assessments my teachers have chosen, but have never really understood why. I now realize that my teachers did not match their assessments to what they taught. Assessment is a part of the *learning process* and therefore all assessment should be intertwined with both targeted learning and instruction." (6)

". . .instructional methods and goals must coincide with what is assessed and the way it is assessed. For example, you cannot teach conceptual knowledge to students and then assessment them on related procedural knowledge." (17)

"My main focus in the unit was based on using supportive evidence and analyzing text. However traditional item assessment doesn't really allow for these objectives to be assessed. Essay writing or other forms of performance assessment would be more valid." (13)

"One of [my first essay questions] was not connected to my objectives. It related to my unit content, but it was simply a question that I <u>wanted</u> to ask, rather than one that I could fairly ask students to write about as a major essay topic, given the

nature of my unit objectives and instruction." (22)

"I embraced [writing essay items] as an opportunity to grapple with questions of validity. For example, are essays appropriate in math classes? Since I believe they are, my next question was what sort of essays are appropriate in math classes? And furthermore, I wondered about how one goes about assessing writing in math classes anyway?" (17)

"Being a math major and having a natural inclination to not want to write, except in mathematical terms, I found the essay items particularly difficult. When would I ever use them in math? What I found as I wrote the questions was that I could probably get a better assessment of what my students actually understood by having them write about it. If I just have them work problems, I can't be sure if they understand the concepts or are just using an algorithm." (19)

"For example, if I am trying to understand whether my students are able to build their ideas toward a conclusion, a multiple-choice test on the components of conclusion building offers less valid information than a short essay where students actually build a conclusion from a thesis. The essay shows the students' ability to apply their knowledge in a holistic setting, while the multiple-choice questions test for definition understanding. Again, the question is can they build a conclusion, not do they know the definition of a conclusion's components. Thus, validity is appeased if the assessment offers the best fit between information gathered and the decision to be made." (18)

"If I think I have taught my students how to write with an audience in mind, then my assessment mechanism must be designed to tell me just that. I know that the information obtained by true/false items will be difficult to support in this regard. I will have to devise some kind of written assessment geared to using the presence of an audience if I am . . . to understand whether my students have met this objective." (18)

"In developing the traditional item assignment, I learned how difficult yet possible it is to develop short answer items which remain true to the discipline of history. . . . However, in revising them, I felt I grasped a very basic point - the items need to assess the actual learning of an objective - not merely the recall of it. Thus I realized if my objectives focus on interpretation and understanding of texts, then that is what the assessment item needs to [ask for] - and that it can be done using traditional items." (20)

"In developing my rating scale, I had to think about constant, observable behaviors that clearly demonstrate that students are engaged in scientific inquiry. . . . I found it challenging to try to think of a number of behaviors that could be applied to any scientific inquiry at any time. . ." (6)

Finally, they had to address the degree to which their scoring rules for essays and performances related to the targeted objectives; many found their scoring rules were far afield (Validity Dimension 1).

"My first rating scale was way off target. The items I wrote did not match my objectives and many of them were inferences about behaviors rather than the behaviors themselves. . . . I thought long and hard about specific behaviors that would demonstrate that students are thoughtfully engaged in scientific inquiry. After writing down each behavior, I tried to imagine my students actually demonstrating this behavior. I then developed a rating scale for scientific inquiry based on these behaviors which was much closer to the targeted learning for this assessment." (6)

"Prior to developing this checklist, I had never considered which particular behaviors demonstrate that a student is comfortable participating in discussion; I merely assumed that it would be clear if a student were UNcomfortable. Drafting this checklist forced me to consider the different components of speaking effectively in a classroom situation. Being aware of the contributing factors will help me give

students much more effective feedback regarding areas in which they need to improve." (10)

"I have always planned for essay items to be a major part of my curriculum. Yet, having to come up with criteria before I graded them, focused me more toward understanding what it was that I wanted my students to do with the items. Basically, my writing progression went like this: I wrote the directions, then I wrote the question, then I wrote the model answer. This led me to rewrite the directions. Then I wrote the criteria and rating scale. This forced me to rewrite the directions once more. It was only then, when the whole thing was complete, that I was truly satisfied that my students were going to be assessed validly, reliably, and fairly." (18)

"The whole idea of creating a good, tight, specific scoring rubric began to make sense to me. . . I realized that is was imperative that I knew and my students knew what I was looking for before I began the assess the essays. This way I am much more focused on the targeted learning objectives and the students are fully aware of what is expected of them." (13)

Grading Policy

The final component of the plan was the grading policy. We required students to use the assessment ideas derived from their unit plans to write a grading policy that could apply to the entire quarter or trimester they had described. The grading policy document was in the form of a handout for students and parents. It explained to students what types of assessments would contribute to their grades (e.g., essays, reports, projects, tests, homework, daily seatwork, etc.) and why this work was important to their learning, what weight would be given to each type of assessment, and how they would go about summarizing across performances to assign a grade (grades based on relative performances or on absolute standards).

In developing this component, students had to think about the role of grading within the validity framework. Students had to grapple with grading issues raised by their experiences both as

students and in their field placements - issues such as how much weight to give to attendance, timeliness, oral participation, and attitude when making judgments about their students' learning of the targeted objectives. By validity standards, these variables would be considered sources of *irrelevant variance* that lead to invalid inferences about student learning (Validity Dimension 1). Because they had been thinking about the relationship between assessments and their disciplines throughout the planning process, it became more difficult for them to accept and adopt inappropriate practices (Stiggins, 1989).

In addition, through readings about the influences of grading practices on motivation and self-esteem, they were forced to question assumptions often made about the motivating power of grades (Covington & Beery, 1976; Canady & Hotchkiss, 1989) and to consider the potential consequences of various ethical and unethical grading practices (Validity Dimension 5). Many students indicated that in being forced to think about the relative weight of each aspect of the grade, they had to look again at the discipline to decide which sources of evidence were best and most important in making judgments about their students' learning (Validity Dimension 1).

> "I learned to represent my beliefs of important parts of my curriculum in grading. . . . The grading policy forced me to reflect on what was valid and fair to grade and what should be taught without being graded . . . This component helped me to see that the grading policy is one of the major ways teachers communicate their learning objectives and goals to parents." (5)
>
> "It is not fair to students to put the bulk of their grade on non-essential performances - that is, they should be graded on the most important performances the heaviest. Also, they should be graded on a variety of items/categories to ensure valid data about the students' performances." (4)
>
> "Since students construct meaning out of how teachers value aspects of an assignment or test, assessment can adversely effect student learning if non-targeted learnings have an undue weight in a student's grade. . . . By designing a grading policy I thought about how to communicate course expectations to students (and

parents) in a way that would facilitate student understanding of course objectives and assignments from the first day of class. I thought of a grading policy as the framework for the class. . ." (1)

"This component reinforced how easy it is for validity of assessment to be in question. Although I tried to develop a grading policy that reflected my teaching goals, there are activities that I eschewed purposefully because I didn't think I could grade them fairly or even teach them (creativity). Yet I do hope to provide opportunities for such activities in my classroom because I consider them important. Yet what does it say to a kid about importance of a subject if it isn't included in a grade?" (23)

The grading policy also gave students an opportunity to think about issues of reliability in assessment. If they had taken the time to create individual assessments that would provide valid information about student learning, then many such assessments were more likely to yield grade summaries that could be trusted as reliable indicators of student learning.

"The question at the heart of reliability is whether the teacher has enough information to make a judgment. Thus a multiplicity and diversity of assessments across time provides greater reliability than a single-shot one dimensional assessment does. Developing a comprehensive grading policy which embraced a number of assessments helped me see how to structure systematic assessments over time so as to increase reliability." (20)

"The best way to be on guard against bias is to use multiple assessment vehicles and acknowledge that bias will always exist. . . . When constructing the unit plan I built in different assessment vehicles. Discussions, essay papers, library research, and conducting surveys and interviews, were all included. Additionally, when I developed my grading policy, I made provisions for individual student contracts. These contracts can be used to offer additional assessment vehicles to students who need to approach the objectives from a different way." (12)

Reflection and Self-evaluation

Once students had completed all of the components of the assessment portfolio, they were given an opportunity to reflect on each piece and discuss (a) what they saw as the strengths of the given component, (b) whether they would change the component now that they had completed all subsequent components and if so, how, and (c) what the component had taught them about the relationship between assessment, instruction, and learning targets. This step gave students a model of "reflective practice" (Schon, 1987) in that good teachers will revise instructional and assessment methods after using them with one or more groups of students.

Students then wrote a self-evaluation of their learnings in the course (as previously described). This final self-evaluation proved to be an excellent tool for the course instructors to use to assess students culminating understandings of validity and reliability. Because students were engaged in an experience that was contextual and where all parts were interrelated, learning was a course-long experience for many. The excerpts given thus far demonstrate that most students had a grasp of the various aspects of validity and their implications for teachers. These students could both discuss and use these concepts in designing classroom assessments.

However, some students were technically able to do the assignments without deep understanding of the concepts. The self-assessment elicited the surface nature of their understandings. These students seemed unable to give specific examples to elaborate on their understandings, and seemed to be merely parroting the instructor's definitions.

> "Validity of assessments refers to how valid the information is to make a correct decision. . . . On essay items, performance assessment, and traditional items in the portfolio, I had to make sure the questions on those items are measuring what I want to measure." (27)
>
> "In designing my specific unit for this class I have learned a great deal about the concepts of validity and reliability. By far the most useful aspect of these concepts are their relationship to one another, and to the learning objectives for the particular unit of study. As far as the concept of validity is concerned, what I understand it to

mean (in relation to assessment) is the degree of accuracy in terms of indicating student knowledge. Referring again to my development of essay questions for this unit, I found myself initially focusing on questions which in no way reflected my learning objectives, and were therefore not valid." (26)

"An assessment is valid when it measures what it claims to measure. The validity of an assessment can be present in degrees to measure to what extent this assessment information will help to make appropriate decisions about students or instruction. The feedback I received on the different components of my portfolio was extremely helpful to me in deriving a valid method of assessment. The components helped me to see exactly what I was missing and what I needed to include to make my assessments as valid as possible." (25)

"The term bias is used to represent when the teacher has used information or data to assess students that may have been favorable or unfavorable to a culture, gender, race, or class. It is related to fairness in that a fair assessment would not use any biased data, formal or informal, to determine knowledge. My forms of assessment must only test my discipline, not how well my students understand the assessment linguistically, culturally, or socially. In reflecting back on this issue in my portfolio, the only examples of humans I used were based on two men and one woman. I don't think any of my forms of assessment were extremely biased but they all were slightly biased because it is impossible to create a completely unbiased form of assessment." (24)

Comparative Studies of the Traditional Assessment Course and the Portfolio Course

Several studies were conducted in an effort to compare the outcomes of the traditional version of the assessment course with those of the portfolio version of the course. Study 1 entailed a comparison of course evaluations across teaching faculty for the two versions of the course. Study 2 involved analyses of data from surveys sent to teacher education students in the quarter following their enrollment in the assessment course - the time during which they were student

teaching full time. Study 3 entailed evaluations of relevant components of an exit survey given to all students exiting from the teacher education program. Each of these studies is described more fully below.

Study 1

Subjects

Toward the end of each quarter, students are administered a course evaluation form. Course evaluations are required for every course for assistant professors and at least once a year for senior faculty. Student participation is voluntary, however, most students complete the form. Results of the course evaluation are not given to the instructor until after grades are submitted. Data were requested for each quarter from the summer quarter of 1988 through the spring quarter of 1994. Date representing 12 quarters of the traditional version of the course and 12 quarters of the portfolio course were available. The number of respondents from the traditional course ranged from 15 to 55 with a mean of 32.25. The number of respondents from the portfolio course ranged from 17 to 74 with a mean of 32.58. Academic ranks for the instructors in the traditional course ranged from teaching assistant to full professor. Academic ranks for the instructors in the portfolio course ranged from teaching assistant to assistant professor. There were 8 different instructors for the traditional course and 3 different instructors for the portfolio course.

Measure

In this study, mean item data from course evaluations for each quarters were obtained from the Office of Educational Assessment at the university. These data were coded by quarter, year, rank of instructor, course evaluation form, and type of assessment course. Course evaluation forms were compared for common items and only those items common to all evaluation forms were evaluated. Items on the forms are given in Table 3. Each item was rated on a 6 point scale. "Excellent" was coded as 5, "very good" was coded as 4; "good" was coded as 3; "fair" was coded as 2; "poor" was coded as 1; "very poor" was coded as 0.

- insert Table 1 about here -

Results

Mean item scores were averaged across classes within each type of course. Only those items specifically related to the content of the course and the relevance of the course were included in the analyses. Two analyses were performed on a selected set of the items. In the first analysis, Item 1, course as a whole, and Item 2, course content, from the "general evaluation" section and Item 3, amount you learned in the course, and Item 4, relevance and usefulness of course content, from the "information to other students" section were summed to obtain an overall score for the *general content* of the course. In the second analysis, Item 4 from the "information to other students" section, *relevance and usefulness*, was analyzed alone. Two t-tests were performed to compare mean ratings for these data. The results of these tests are given in Tables 2 and 3.

- insert Tables 2 and 3 about here -

As can be seen from these data, there were significant differences between students perceptions of the *general content* of the course ($t = -5.85$, $p < .001$) and between students perceptions of the *relevance and usefulness of the course* ($t=7.00$, $p < .001$). Students in the portfolio course clearly saw the course as more relevant to their needs and rated the content of the course between very good and excellent

One possible explanation for these differences could have been the differences in instructors. However, even instructors who received high ratings for *instructor's effectiveness* (Item 4 of the general evaluation section) received lower ratings on *relevance and usefulness of course content*. The two instructors from the traditional course with the highest ratings for *instructor's effectiveness* had mean ratings of 4.38 and 4.25. Their mean ratings for *relevance and usefulness* were 3.52 and 3.83 respectively. The mean ratings for these instructors for *course content* were 3.90 and 3.64 respectively. In addition, the instructor of the portfolio course with the lowest mean ratings for *instructor's effectiveness* (2.41) had a higher mean ratings for *relevance and usefulness* (3.52) and for *course content* (3.31). This suggests that the students' perceptions of the effectiveness of an instructor was somewhat independent of whether they saw the content of the assessment course as relevant to their needs.

Study 2

Subjects

In the second study, students from six different quarters were asked to volunteer to be part of a survey during the following quarter of their program. Most of the students were engaged in full-time student teaching during the quarter following the one in which they took the assessment course. Two classes of students (N = 112) who had taken the traditional course during the summer of 1992 were surveyed. Twenty-one percent (n = 23) of these students completed and returned the surveys. Five classes of students (N = 195) who had taken the portfolio version of the course between the summer of 1991 and the autumn of 1992 were surveyed. Twenty-five percent (n = 50) of those enrolled completed and returned the surveys.

Measure

A questionnaire was developed regarding a number of assessment and programmatic issues. The items relevant to these studies are given in Table 4.

- insert Table 4 about here -

Responses to items 4, 6, and 7 (the influence of assessment, validity issues, and reliability issues respectively) were coded by two raters. Coding was based on the degree to which the students' responses showed understanding of general assessment concepts. Table 5 provides the scheme used to code student responses.

- insert Table 5 about here -

Interjudge agreement for the ratings on these items ranged from 71.2 to 98.6. All discrepancies were resolved through a discussion between the raters. Once discrepancies were resolved, counts for each code in each group for each item were compared using a chi square statistic. For the traditional group, 26% indicated that the course had had no effect on their teaching. For the portfolio group, 1% indicated that the course had had no effect on their teaching. The counts for each group for the remaining codes of items 4, 6, and 7 are given in Tables 6 through 8. As can be seen for item 4, a significantly greater percent of students from the portfolio course showed a clear understanding of the appropriate uses of assessment than did students in the

traditional course ($\chi^2_{(3)} = 17.75, p < .001$). For item 6, a significantly greater percent of students from the portfolio group gave good examples of validity issues than did students from the traditional course ($\chi^2_{(3)} = 14.45, p < .003$). For item 7, a significantly greater percent of students from the portfolio group gave good examples of reliability issues than did students from the traditional course (($\chi^2_{(3)} = 9.51, p < .03$), however, a fairly large proportion of both groups gave no examples at all (65% of the traditional course students and 42% of the portfolio course students). In addition, a fairly large percent of the students in the portfolio group (32%) received a score of 1 for this item, indicating that while the portfolio group may have been better prepared to address issues related to reliability than were the students in the traditional version of the course, they were not sufficiently prepared regarding issues of reliability.

- insert Tables 6 through 8 about here -

Study 3

Subjects

As part of the ongoing evaluation process of the Teacher Education Program, exit surveys were administered in the last quarter of the program to all students. We obtained 129 of these surveys from three years just prior to the change (1989-91) in the assessment course and 151 from two years after the change (1992 and 1994). In the summer of 1992 an outside instructor taught a traditional course. Since it was not possible to tell which students finishing in 1993 had taken the revised course, data from this year were not used.

Measures

Exit surveys asked a variety of questions about students' experiences in the Teacher Education Program, including both course work and field work. First, a set of items asked students to rate how well the program as a whole had prepared them in a number of areas corresponding to the state requirements for teacher education programs. One of these items was "How well has this program prepared you to evaluate student work," which students rated on a scale from 1 ("not at all prepared") to 5 ("thoroughly prepared").

A set of open-ended questions asked students to comment on various program aspects.

The first question asked for comments about courses in the program. Comments specifically directed at the assessment course, and related to value or worth of the course or its content were coded (0) if they suggested eliminating the course altogether; (1) if they stated the course was worthless, not valuable, not useful for teachers; and (2) if they stated the course was valuable, applicable or useful. General comments (not referring to value) were coded (1) negative or (2) positive. A second item asked students to list aspects of the teacher education program that were particularly valuable or worthwhile; we counted the number of students listing the assessment course here. A third item asked what important material was left out or not sufficiently covered; we counted any mention of an assessment-related topic (e.g., setting up grade books, portfolios, informal observation). Finally, negative comments regarding the work load were counted.

Results

Ratings of how well students thought the program prepared them to do assessment were compared across courses using a one-way ANOVA with unique sums of squares. Students who took the new course rated the teacher education program as preparing them more thoroughly to do assessment ($\underline{M}$ = 4.09, $\underline{SD}$ = 0.87) than did students who took the traditional course ($\underline{M}$ = 3.16, $\underline{SD}$ = 1.06; $\underline{F}_{(1, 280)}$= 65.44, $\underline{p} < .001$). The difference in assessment course accounted for approximately 19% of the variance in ratings for the program as a whole (including er courses and three to four quarters student teaching).

Frequency of responses for each open-ended item appear in Table 9. In general, the comments were more positive for the revised course, though not uniformly so. Typical comments for the traditional course included "308 was a useless class. Testing and evaluation are essential, but I learned almost nothing in this class" and "Did not relate to the real world". Typical comments for the revised course included "308 provided me with the information that I considered most valuable in my field experience" and "08 was the most valuable class overall for my teaching." Eight students (5.2%) stated that the work load in the revised course was excessive, while none of the students taking the traditional course did so.

Each comment was coded into only one category, but some students mentioned the

assessment course in more than one way. Therefore a new variable was created by counting the number of students in each group who had responded in some way that the assessment course was valuable and the number of students who had indicated that the course was not valuable. These results are given in Table 10. Students who had taken the revised course were much more likely to rate it as valuable, while those taking the traditional course were more likely to see the course as not valuable ($\chi^2 = 61.8$, $p < .001$).

- insert Tables 9 and 10 about here -

Discussion

As can be seen from the work of the students and the studies comparing students in the portfolio version of the course with students in the traditional version of the course, the portfolio course was more successful in helping students learn and retain assessment concepts. Each component of the assessment portfolio provided an opportunity for students to address one or more of the dimensions of validity highlighted in this paper. The focus on validity guided student learning from the initial course description and concomitant goals and objectives which helped students develop clearer definitions of their disciplines for themselves, to the unit assessments which helped students explore all five dimensions of validity, to the grading policy which helped them address issues of multiple sources of evidence, appropriateness of evidence, and potential consequences of assessment interpretations and use.

Ultimately, the power of this course may lie in the fact that it was a model of the concepts students were learning: the portfolio demonstrated interdependence of instruction and assessment in influencing student learning, the components of the portfolio were designed to teach and assess the students' learning related to the targeted course objectives, and multiple sources of evidence were used to gather information about student learning. Former students (now teachers) tell us they can no longer think about planning for instruction without thinking about assessment at the same time. The students in this cohort saw that a clear relationship between learning targets, instruction and assessment was fundamental to student learning.

"I have learned that assessment, when properly designed and implemented, can

actually be part of instruction. . . . I had my students doing research on historical accounts, similar events in modern times, comparing the two events, and writing analytical essays in cooperative groups. . . . The essays were both methods of instruction and assessment vehicles." (12)

"It is fair to allow students to know what is expected of them during a lesson or during a unit by giving them the criteria for assessment. Not only does this knowledge increase student comfort level during a learning episode, but this knowledge scaffolds student learning within targeted learning objectives." (1)

"Since the task of assessing was so difficult, I was forced to answer the question at each point of whether the learning was so necessary to an understanding of the subject matter. Observing that assessment of a meaningless bit of learning was a waste of time for me sprouted the idea that perhaps it was a waste of time for my students, as well. If something is not worth assessing, at however informal a level, it isn't worth teaching or learning, either." (3)

"I had always considered assessment to take place outside and after instruction. Now I see assessment as another valuable step in the education process. When the fit between objective and assessment is harmonious, then the tool used for assessment becomes another step in the instruction of the objective. . .If I design reliable, valid, and fair assessments, my class time will be spent in constant pursuit of education. . .Furthermore, true assessments can act to bridge their learning from the class towards real world application of their knowledge, as opposed to mark sense application. The nature of the power of assessment is still the same, it is just that I now have more helpful and useful ways of harnessing that power toward educational ends." (18)

"[A] description of how I will assess an assignment provides scaffolding for the difficult assignment while still providing students with room for individual student creativity. . . . [T]elling students how they will be assessed helps to clarify what is

expected of students [and] helps students to focus on the targeted learnings during instructional time." (1)

The focus on validity also helped these students recognize the power of assessment in their classrooms.

"Without a strong relationship [among objectives, activities, and assessment], my classes will be doomed for failure. . . . I have to admit that when I planned things for my student teaching that I really did not take into account too much of what I wanted my students to learn or really anything about assessment at all. . . . I now see that in order to meet my overall goal of augmenting student learning, that I have to design assessments that will dominate my class in a way that supports student learning rather than tears it down." (15)

"If students do not see a clear relationship between learning objectives, teaching, and assessment, then assessments will tend to discourage students from being engaged in class." (17)

"Students pick up on what is important to a teacher by what is assessed and how it is assessed. If assessment choices support student learning of targeted objectives then students know that student learning is the bottom line of success in the course." (1)

"The assessments have to come from the goals and objectives. If they don't, the students learn that what they do in class does not really matter. In effect, they don't learn the objectives because the assessment does not reflect how much they learned. If the assessments are good, come from the objectives, and relate to the discipline, students see that what they learn is validated by a score or grade. What they do in class makes sense, and this helps to encourage further growth in the discipline." (8)

Through the words of our students, we have described the impact of one course in assessment on the thinking of a group of future teachers. Of course, the students will struggle with implementing these and other "best practices" in their work as full-time teachers. The fact that they

can discuss specifically the nature of validity as it relates to particular teaching problems in particular units of study, however, suggests that their understanding of validity is profound. Exit survey data, course evaluation data, and follow-up survey data show that students valued the course and saw a clearer connection between the assessment course and their roles as teachers. In fact, the most frequent criticism of the course was that is was too short. The focus on validity helped students include assessment into their frameworks for teaching and learning in ways that should enable them to develop and implement valid assessment in their own classrooms - assessments that support student learning.

> "This class has helped me to understand the nature of learning better by realizing what a good, valid assessment is. . .I come away knowing that assessments can be powerful learning tools as well as positive experiences if handled with care." (16)

Table 1

Course evaluation items common across all evaluation forms

Section	Item	Stem
1: General Evaluation	1	Course as a whole
	2	Course content
	3	Instructor's contribution to the course
	4	Instructor's effectiveness in teaching the subject matter
2: Feedback to Instructor	1	Course organization
	3	Explanations by instructor
	4	Instructor's ability to present alternative explanations
	5	Instructor's use of examples and illustrations
	7	Student confidence in instructor's knowledge
	8	Instructor's enthusiasm
	11	Availability of extra help when needed
3: Information to Other Students	1	Use of class time
	2	Instructor's interest in whether students learned
	3	Amount you learned in the course
	4	Relevance and usefulness of course content
	5	Evaluative and grading techniques (tests, papers, projects)
	6	Reasonableness of assigned work
	7	Clarity of student responsibilities

Table 2

T-test results for: *general content* score

	Number of Cases	Mean	Standard Deviation	t Value	Degrees of Freedom	2-Tail Prob.
Traditional Course	12	12.0925	2.0368	-5.85	22	.000
Portfolio Course	12	16.4842	1.6187			

Table 3

T-test results for: *relevance and usefulness* mean score

	Number of Cases	Mean	Standard Deviation	t Value	Degrees of Freedom	2-Tail Prob.
Traditional Course	12	2.9233	.56545	-7.00	22	.000
Portfolio Course	12	4.2958	.37558			

Table 4

Post-course survey items related to assessment concepts

Item	Stem
4	Thinking back on (the course) have any ideas or other aspects of the course influenced your teaching? If so, what part of (the course) has influenced your teaching the most? How has this influenced your teaching?
6	Have you wrestled with any validity issues in your field placement this quarter? If so please describe one such issue.
7	Have you wrestled with any reliability issues in your field placement this quarter? If so please describe one such issue.

Table 5

Coding scheme for relevant items of the post-course survey

4 Influence of course on teaching

Code 1 1 = yes 2 = no

Code 2

2 = shows clear, unambiguous understanding of appropriate uses of assessment

1 = • shows partial understanding of appropriate uses of assessment

• describes delivery of instruction; may have assessment links

• uses assessment terms without examples

0 = shows little or no understanding of appropriate uses of assessment in instruction

NS = not scorable (off task or omitted)

6 Validity issues

2 = gives good example of validity issue

1 = • possible example of validity issue, somewhat unclear

• may confuse validity with reliability

0 = gives example that is neither reliability nor validity

NS = not scorable (off task or omitted)

7 Reliability issues

2 = gives good example of reliability issue

1 = • possible example of reliability issue, somewhat unclear

• may confuse validity with reliability

0 = gives example that is neither reliability nor validity

NS = not scorable (off task or omitted)

Table 6

Chi square results comparing post-course survey items for students from the traditional and portfolio courses for item 4: Influence of assessment course

	Number of Cases	Percent for Each Code 2	1	0	NS
Portfolio Course	50	35 (70%)	13 (26%)	1 (2%)	1 (2%)
Traditional Course	23	10 (43%)	3 (13%)	2 (9%)	8 (35%)

Table 7

Chi square results comparing post-course survey items for students from the traditional and portfolio courses for item 6: Validity issues wrestled with

	Number of Cases	Percent for Each Code 2	1	0	NS
Portfolio Course	50	33 (66%)	5 (10%)	1 (2%)	11 (22%)
Traditional Course	23	5 (22%)	3 (13%)	3 (13%)	12 (52%)

Table 7

Chi square results comparing post-course survey items for students from the traditional and portfolio courses for item 7: Reliability issues wrestled with

	Number of Cases	Percent for Each Code 2	1	0	NS
Portfolio Course	50	12 (24%)	16 (32%)	1 (2%)	21 (42%)
Traditional Course	23	2 (9%)	3 (13%)	3 (13%)	15 (65%)

Table 9

Frequency of responses to each item.

	Number	Comments (Value)		
Course	of Cases	Valuable	Not Valuable	Eliminate Course
Traditional	129	1	17	9
Revised	154	19	2	0

	Number	Comments (General)		
Course	of Cases	Positive	Negative	Negative Work Load
Traditional	129	0	9	0
Revised	154	22	4	8

		What aspects of program were	
	Number	Particularly	Not Sufficiently
Course	of Cases	Valuable	Covered
Traditional	129	1	4
Revised	154	28	3

Table 10

Number indicating the course was valuable or not valuable.

	Valuable	Not Valuable
Traditional	1 (0.8%)	26 (20.2%)
New	45 (29.2%)	2 (1.3%)

References

Airasian, P. (1991). Perspectives on measurement instruction for pre-service teachers, *Educational Measurement: Issues and Practice, 10* (1) 13-16, 26.

Airasian, P. (1993). *Classroom Assessment*, Second Edition. New York: McGraw-Hill.

Canady, R. L., & Hotchkiss, P. R. (1989). It's a good score! Just a bad grade. *Phi Delta Kappan, 71* (1), 68-71.

Covington, M. V., & Beery, R. G. (1976). *Self-worth and school learning*. New York: Holt, Rinehart, p. 42-63, 77-87.

Crooks, T. J. (1989). The impact of classroom evaluation practices on students. *Review of Educational Research, 58* (4), 438-481.

Grossman, P. L., Wilson, S. M., & Schulman, L. S. (1989). Teachers of substance: Subject matter knowledge for teaching. In M. C. Reynolds (Ed.), *Knowledge base for the beginning teacher.* New York: Pergamom.

Gulickson, A. R. (1986). Teacher education and teacher-perceived needs in educational measurement and evaluation. *Journal of Educational Measurement, 23* (4), 347-354.

Linn, R. L. (1990). Essentials of student assessment: From accountability to instructional aid. *Teachers College Record, 91* (3), 422-436.

Messick, S. (1989). Validity. In Educational Measurement, Robert Linn (Ed.), Schafer, W. D. (1991). Essential assessment skills in professional education of teachers. *Educational Measurement: Issues and Practice, 10* (1), 3-6, 12.

Resnick, L. B. & Resnick, D. P. (1991). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction.* Boston: Kluwer.

Schafer, W. D., & Lissitz, R. W. (1987). Measurement training for school personnel: Recommendations and reality. *Journal of Teacher Education, 38* (3), 57-63.

Schon, D. A. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions.* SF: The Jossey-Bass Higher Education Series.

Schwab, J. J. (1978). *Science, curriculum, and liberal education.* Chicago: University of Chicago Press.

Stiggins, R. J. (1988). Revitalizing classroom assessment: The highest instructional priority. *Phi Delta Kappan, 70* (5), 363-368.

Stiggins, R. J. (1989). Inside high school classroom grading practices: Building a research agenda. *Educational Measurement: Issues and Practice, 8* (2), 5-14.

Stiggins, R. J. (1991). Relevant training for teachers in classroom assessment. *Educational Measurement: Issues and Practice, 10* (1), 7-12.

Stiggins, R. J. (1994). *Student centered classroom assessment..* New York: Merrill, an imprint of Macmillan College Publishing Company.

Stiggins, R. J., & Bridgeford, N. J. (1988). The ecology of classroom assessment. *Journal of Educational Measurement, 22* (4), 271-286.

Stiggins, R. J., & Faires-Conklin, N. (1988). Teacher training in assessment. Portland, OR: Northwest Regional Educational Laboratory.

Stiggins, R. J., & Faires-Conklin, N. (1992). *In teachers' hands: Investigating the practices of classroom assessment..* Albany, NY: SUNY Press.

Stiggins, R. J., Faires-Conklin, N., & Bridgeford, N. J. (1986). Classroom assessment: A key to effective education. *Educational Measurement: Issues and Practice, 5* (2), 5-17.

Valencia, S. (1990). A portfolio approach to classroom reading assessment: The whys, whats, and hows. *Reading Teacher, 43* (4), 338-340.

Wise, S. L., Lukin, L. E., & Roos, L. L. (1991). Teacher beliefs about training in testing and measurement. *Journal of Teacher Education, 42* (1), 37-42.

Wolf, D. P. (1991). Assessment as an episode of learning. In R. Bennett and W. Ward (Eds.), *Construction versus choice in cognitive measurement.* Hillsdale, NJ: Lawrence Erlbaum Associates.